Individuazione delle tipologie di pendolari nel Friuli Venezia Giulia

ADRIANA MONTE, GABRIELLA SCHOIER

DEAMS, Università di Trieste

1. Introduzione

Nella società moderna la mobilità degli individui ha assunto notevole importanza. In particolare in questo lavoro si discute del ruolo che il pendolarismo, per motivi di studio e di lavoro, ha nei sistemi socio-economici (Goodman, 2013). L'analisi e la modellizzazione di *networks* di pendolari è diventata cruciale per lo studio delle dinamiche delle diverse aree geografiche; inoltre ciò risulta fondamentale per decisioni in ambito politico-sociale; in quest'ottica l'analisi del pendolarismo e i suoi effetti sul mercato del lavoro, sulla pianificazione territoriale e delle infrastrutture può essere vista come parte della modellizzazione spaziale in un contesto di Data Mining Spaziale (Han et al., 2001).

Per quanto riguarda l'Italia alcune domande sulla mobilità per motivi di studio e lavoro sono state introdotte nel 1971 nell'ambito del Censimento della popolazione. Con il Censimento del 1981 è iniziata la costruzione delle matrici degli spostamenti per la mobilità tra comuni in relazione ai flussi casa lavoro/studio. Nel 2011 considerevoli innovazioni sono state introdotte; tra queste l'utilizzo di indagini campionarie per ottenere stime per alcune variabili di carattere socio-economico che nei precedenti Censimenti venivano rilevate su tutta la popolazione. Due tipi di questionario sono stati utilizzati: il questionario in forma long (L) che è stato distribuito a tutte le famiglie nei comuni con meno di 20000 abitanti e ad un campione di famiglie nei comuni capoluogo di provincia o con almeno 20000 abitanti e il questionario in forma short (S) distribuito alle restanti famiglie.

In questo lavoro viene utilizzata la matrice del pendolarismo derivata dal Censimento 2011 per individuare le tipologie di pendolari della regione Friuli Venezia Giulia. Ciascun *record* della matrice rappresenta uno strato di pendolari descritto dalle variabili presenti nella matrice stessa.

La metodologia di analisi utilizzata nel lavoro è la *cluster analysis*, in particolare la *two step cluster analysis* che permette considerare database di diverse dimensioni sulle unità del quale si possono rilevare sia variabili qualitative che quantitative per individuare cluster di pendolari con caratteristiche in comune.

2. La mobilità nei comuni del Friuli Venezia Giulia per studio e lavoro

Prima di procedere alla descrizione della metodologia e all'illustrazione dei risultati dell'analisi, si descrive un quadro sintetico del pendolarismo in Friuli Venezia Giulia, come risulta dai dati censuari. In questa regione più del 64% della popolazione residente vive in comuni di medio piccole dimensioni (con meno di 20000 abitanti), mentre il resto dei residenti è distribuito tra le città capoluogo (Gorizia, Pordenone, Trieste e Udine) e altri due centri urbani (Monfalcone e Sacile) che superano i 20000 abitanti1. Poco più del 50% della popolazione dichiara di essere pendolare per motivi di studio o lavoro (si veda Tavola 1) e l'incidenza di questo pendolarismo è più elevata nelle province di Pordenone e Udine, che presentavano al 2011 un più elevato tasso di occupazione e un indice di vecchiaia minore (Stassi et al., 2013).

Tavola 1 – Popolazione residente e pendolari in Friuli Venezia Giulia al Censimento 2011

	Popolazione residente (R)	Pendolari residenti (P)				P/R %
		Totale	in comunità	in famiglia		
				per studio	per lavoro	
Gorizia	140143	68356	37	20008	48311	48,7
Pordenone	310811	163221	93	47426	115702	52,5
Trieste	232601	114967	147	33319	81501	49,4
Udine	535430	270895	169	77831	192895	50,5
Total FVG	1218985	617439	446	178584	438409	50,6

Fonte: Istat – Censimento 2011

Sulla base dei dati del Censimento 2011, la modalità di spostamento più utilizzata per recarsi al luogo di studio o lavoro è l'automobile, analogamente a quanto risultava dai precedenti Censimenti (si veda la Tavola 2). Caratteristiche diverse rispetto al resto della regione presenta la mobilità nel comune di Trieste, che è la città di maggiori dimensioni della regione stessa (circa 200000 abitanti); anche se l'automobile continua ad essere la modalità di spostamento più utilizzata in questa città e nella sua provincia, la distribuzione dei pendolari secondo le diverse modalità di trasporto è diversa rispetto a quelle delle altre province. Questo è dovuto non solo alla dimensione della città, ma anche alla sua conformazione geografica.

Di solito la modalità utilizzata in regione dai pendolari per spostarsi dipende dal motivo del pendolarismo; coloro che si spostano giornalmente per motivi di lavoro usano prevalentemente l'automobile come conducente (70,3%), mentre chi si sposta per motivi di studio usa altre modalità, il 33,2% usa l'autobus, il 37% si sposta in automobile (come passeggero) e il 16,5% va a piedi.

Per quanto riguarda il tempo impiegato la maggior parte dei pendolari impiega non più di 15 minuti per arrivare al luogo di studio o di lavoro come si vede dalla Tavola 3; inoltre l'87,3% dei pendolari per motivi di lavoro e l'83,2% dei pendolari per motivi di studio impiegano al massimo 30 minuti.

1 La dimensione demografica dei comuni fa riferimento alle risultanze anagrafiche del 31 dicembre 2010, utilizzate per costruire le liste censuarie e definire la dimensione dei comuni. In seguito alle rilevazioni censuarie il comune di Sacile risultava avere una popolazione minore di 20000 abitanti.

Tavola 2 – Distribuzione percentuale dei pendolari secondo modalità di spostamento in Friuli Venezia Giulia ai Censimenti 1991, 2001, 2011

	1991	2001	2011
Treno, tram	2,96	1,9	1,9
Autobus	18,5	13,8	13,7
Auto (come conducente)	43,9	51,1	50,8
Auto (come passeggero)	9,1	12,1	13,3
Bicicletta	4,8	5,2	3,5
Altro	20,7	15,9	16,7

Fonte: Istat – Censimenti 1991, 2001 e 2011

Tavola 3: Distribuzione percentuale dei pendolari secondo tempo impiegato per arrivare al luogo di lavoro o di studio in Friuli Venezia Giulia al Censimento 2011

	Lavoro %	Studio %
Fino a 15minuti	55,6	62,6
16 – 30	31,7	21,6
31 – 45	7,0	7,1
46 – 60	3,3	5,0
Oltre 60 minuti	2,4	3,7
Totale	100,0	100,0

Fonte: Istat – Censimento 2011

3. La metodologia: generalità sulla *cluster analysis*

Esistono diversi metodi statistici per la classificazione delle unità in gruppi omogenei; essi possono essere suddivisi in due grandi categorie: *supervised classification* e *unsupervised classification*. Nel primo caso si hanno a priori n unità osservate che appartengono a due o più popolazioni differenti e di ognuna si conoscono i valori delle p variabili considerate. Lo scopo di questo tipo di analisi è l'assegnazione di ulteriori unità alla popolazione di appartenenza, minimizzando la probabilità di errore di attribuzione. Al contrario i metodi del tipo *unsupervised classification* sono tipicamente esplorativi e consistono nella ricerca, nelle n osservazioni p–dimensionali, di gruppi di unità tra loro simili, non sapendo a priori se tali gruppi omogenei esistono effettivamente nel dataset (Zani et al., 2007). La classificazione non supervisionata, o *cluster analysis*, ha quindi l'obiettivo di riconoscere dei gruppi che si caratterizzano per un'elevata omogeneità all'interno e per un'elevata eterogeneità tra di essi.

Gli ambiti di applicazione della *cluster analysis* sono molto vari, essa è utilizzabile per: ridurre i dati in forma grafica (per evidenziare le più importanti informazioni rilevate oppure per presentare i risultati di analisi multivariate), generare ipotesi di ricerca (prima di provare un qualsiasi modello di analisi sui dati rilevati è utile individuare le connessioni reali tra le entità e intuire in base a queste i modelli presenti nei dati), individuare gruppi di unità con caratteristiche distintive che, nell'insieme, facciano

percepire la fisionomia del sistema sociale osservato, costruire sistemi di classificazione automatica, stratificare popolazioni da sottoporre a campionamento.

Gli algoritmi di *clustering* suddividono i dati in un certo numero di cluster (o gruppi, sottoinsiemi, categorie). Non esiste una definizione univoca per queste procedure, anche se molti studiosi convergono nel riconoscere un cluster nel caso di omogeneità all'interno di ogni gruppo e di eterogeneità tra i diversi gruppi.

Si supponga di considerare un dataset di $N = \{1, 2, \dots, n\}$ unità e di disporre per ognuna le rilevazioni su p variabili; queste informazioni vengono inserite nella matrice dei dati $\boldsymbol{X}$ di dimensioni (nxp), il *partitional clustering* consiste nel ricercare una partizione di $\boldsymbol{X}$ in K gruppi $(K \leq N)$, $C = \{C_1, \dots, C_k\}$ tali che:

$$C_i \neq 0 \quad \text{per} \quad i = 1, \dots, K;$$

$$\bigcup_{i=1}^{K} C_i = \boldsymbol{X};$$

$$C_i \cap C_j = \emptyset \quad \text{con} \quad i, j = 1, \dots, K, \ \ i \neq j.$$

Come si nota dai vincoli imposti, ogni osservazione appartiene ad un unico cluster tuttavia è possibile allentare questa restrizione e supporre che un'osservazione appartenga a tutti i cluster con un certo grado di appartenenza, $u_{i,j} \in [0,1]$, che rappresenta il coefficiente di appartenenza della j-esima osservazione all'i-esimo cluster in questo caso si parla di *fuzzy clustering* di cui non si tratta nel presente lavoro.

La *cluster analysis* tradizionale consiste in quattro semplici fasi strettamente collegate tra loro; come si vede in Fig.1.1, la procedura può richiedere una serie di tentativi e di ripetizioni dei vari passaggi che vengono di seguito sintetizzati: scelta delle variabili; scelta dell'algoritmo di *clustering*; validazione dei cluster; interpretazione dei risultati.

Esistono diversi algoritmi di *clustering* che possono essere classificati secondo il seguente schema: metodi gerarchici (agglomerativi, divisivi), metodi non gerarchici, metodi basati sull'errore quadratico, metodi basati sui modelli mistura, metodi basati sulla teoria dei grafi, *two step clustering* e altri metodi.

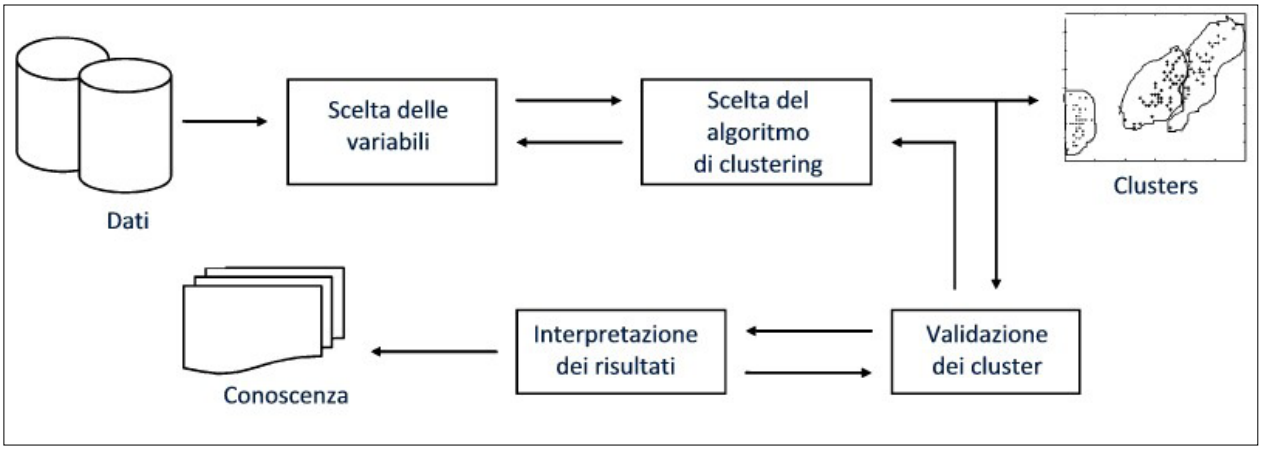

Figura 1 – Le fasi di una procedura di clustering
Fonte: Xu, 2005

Si supponga di partire da un dataset di $N = \{1, 2, \dots, n\}$ unità e di possedere per ognuna le rilevazioni per p variabili. Si consideri la matrice dei dati $\boldsymbol{X}$ di dimensioni (nxp):

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

dove x_{ih} rappresenta l'osservazione della variabile h sulla unità i, con $h = 1, \dots, p;\ \ i = 1, \dots, n$.

Secondo i tradizionali metodi di *clustering*, per individuare dei gruppi di unità omogenei è fondamentale ricavare per ogni coppia di elementi degli indici di prossimità; grazie a questi è possibile raggruppare le n unità in g sottoinsiemi e "ridurre le dimensioni" dello spazio $\mathbb{R}^n$.

Un indice di prossimità tra due generiche unità statistiche u_i e u_j è definito come funzione dei vettori riga della matrice dei dati $\boldsymbol{X}$:

$$IP_{ij} = f\left(\boldsymbol{x}_i', \boldsymbol{x}_j'\right) \qquad i, j = 1, 2, \dots, n.$$

Gli indici di prossimità vengono abitualmente distinti tra indici di dissimilarità (ai quali appartengono le distanze), applicati solitamente a fenomeni quantitativi e indici di similarità, utilizzati per fenomeni qualitativi (Zani et al., 2007).

3.1 I METODI GERARCHICI

I metodi gerarchici aggregativi consentono di ottenere una famiglia di partizioni delle *n* unità statistiche partendo da quella in cui tutte le unità sono distinte, per arrivare a quella in cui tutte sono riunite in un unico gruppo; sono quindi dei metodi che si sviluppano per fasi ordinate, in modo che ad ogni passo vengano uniti i due gruppi (o le due unità nella prima fase) scelti a seconda della misura di distanza (o di similarità) precedentemente definita[2].

Gran parte dei metodi gerarchici partono dalla matrice di distanze $\boldsymbol{D}$ calcolata per le *n* unità statistiche. In questo caso la procedura generale per il raggruppamento delle unità è il seguente.

Fase 1: si individuano nella matrice $\boldsymbol{D}$ le due unità tra loro più simili (in pratica quelle con minor distanza[3]) e si aggregano. Questo è il primo gruppo di unità. Si ottiene una partizione con $(n - 1)$ gruppi, di cui $(n - 2)$ costituiti da singole unità e l'altro formato da due unità.

Fase 2: si ricalcolano le distanze del gruppo ottenuto dagli altri gruppi (alcuni saranno costituiti da una sola unità), ottenendo una nuova matrice delle distanze con dimensioni diminuite di uno.

Fase 3: si individua nella nuova matrice delle distanze la coppia di gruppi (o unità) con minore distanza, unendoli in un solo gruppo.

Fase 4: si ripetono la fase 2 e la fase 3 fino a che tutte le unità sono riunite in un unico cluster.

2 Questa tipologia di metodo viene anche definita *bottom up*, poiché parte dalle singole unità statistiche (dal basso) e procede ad unirle in gruppi sempre più grandi.

3 Se si parte da una matrice di indici di similarità non si guarda alla minore distanza ma alla "maggiore similarità".

La differenza tra i diversi metodi gerarchici consiste solamente nel diverso criterio utilizzato per calcolare la distanza tra due gruppi di unità. Si supponga di avere due cluster C_1e C_2formati rispettivamente da n_1e n_2 unità: sono possibili diverse definizioni di distanza tra i due gruppi, che identificano altrettanti metodi gerarchici. Di seguito si riportano i metodi più noti.

Nel metodo del legame singolo (o del vicino più prossimo) la distanza tra i due gruppi è definita come il minimo delle $n_1 n_2$ distanze tra ciascuna delle unità di un gruppo e ciascuna delle unità dell'altro gruppo: $d(C_1, C_2) = \min(d_{rs})$ con $r \in C_1$, $s \in C_2$.

In pratica con questa definizione, ad ogni passo si valuta la distanza tra due cluster attraverso la distanza dei punti più vicini.

Al contrario, nel metodo del legame completo (o del vicino più lontano) la distanza tra i due gruppi è definita come il massimo delle $n_1 n_2$ distanze tra ciascuna delle unità di un gruppo e ciascuna delle unità dell'altro gruppo:

$$d(C_1, C_2) = \max(d_{rs}) \quad \text{con } r \in C_1 \text{ , } s \in C_2.$$

Con questo metodo tutte le distanza tra le unità di C_1e le unità di C_2 sono minori (o uguali) alla distanza definita dal criterio.

Infine il metodo del legame medio tra i gruppi, secondo il quale la distanza tra due cluster viene definita tramite la media aritmetica delle $n_1 n_2$ distanze tra ciascuna delle unità di un gruppo e ciascuna delle unità dell'altro gruppo:

$$d(C_1, C_2) = \frac{1}{n_1 n_2} \sum_r \sum_s d_{rs} \text{ con } r \in C_1, s \in C_2.$$

I metodi visti fino a questo punto richiedono solamente la conoscenza della matrice delle distanze. Vi sono altri metodi gerarchici che utilizzano anche la matrice dei dati di partenza.

Il metodo del centroide prevede che la distanza tra due cluster C_1e C_2 venga calcolata come la distanza tra due centroidi $\overline{\boldsymbol{x}}_1$e $\overline{\boldsymbol{x}}_2$: $d(C_1, C_2) = d(\overline{\boldsymbol{x}}_1, \overline{\boldsymbol{x}}_2)$.

Il centroide iniziale non è altro che il vettore che contiene i valori medi delle p variabili per le unità incluse nel gruppo. Al passo successivo, il centroide del nuovo cluster potrà essere calcolato come una media aritmetica dei centroidi dei due gruppi iniziali:

$$centroide\ (C_1 \cup C_2) = \frac{\overline{\boldsymbol{x}}_1 n_1\ \overline{\boldsymbol{x}}_2 n_2}{n_1 + n_2}.$$

Il metodo del centroide presenta alcune analogie con il metodo del legame medio: in quest'ultimo si considera la media delle distanze tra le unità dell'uno e dell'altro gruppo, mentre nel metodo del centroide si individua prima il "centro" di ogni gruppo e si misura poi la distanza tra essi.

Nel metodo di Ward (o della minima devianza) non è richiesto il calcolo preliminare di una matrice delle distanze, ma si definisce esplicitamente una funzione obiettivo. Poiché, come già ribadito, lo scopo della classificazione è quello di ottenere gruppi con la maggiore coesione interna, si considera la scomposizione della devianza totale (indicata con la lettera T) delle p variabili in devianza nei gruppi (*Within*, indicata con la lettera W) e devianza tra i gruppi (*Between*, indicata con la lettera B):
se si suppongono g gruppi si definisce

$$T = \sum_{s=1}^{p} \sum_{i=1}^{n} \left(x_{is} - \bar{x}_{s,l}\right)^2;$$

la devianza totale delle p variabili, ottenuta come somma delle devianze delle singole variabili rispetto alla corrispondente media generale $\bar{x}_s$. Questa devianza può essere così scomposta:

$$T = W + B = \sum_{s=1}^{p} \sum_{i=1}^{n_l} \left(\bar{x}_{s,l} - \bar{x}_s\right)^2 + \sum_{s=1}^{p} \sum_{i=1}^{n_l} n_l \left(\bar{x}_{s,l} - \bar{x}_s\right)^2.$$

Il primo termine rappresenta la somma delle devianze di gruppo, il secondo la somma, calcolata su tutte le variabili, delle devianze ponderate delle medie di gruppo rispetto alla corrispondente media generale.

Questa procedura aggrega ad ogni passo i cluster aventi la minima devianza nei gruppi (*Within*), cioè i cluster più omogenei al loro interno.

In generale è possibile rappresentare graficamente le varie partizioni che si ottengono nelle varie fasi di un metodo gerarchico mediante un albero n-dimensionale che viene definito dendrogramma. Questo diagramma ad albero permette di visualizzare i gruppi ottenuti ad ogni stadio dell'operazione di *clustering*.

Negli anni recenti, con la necessità di lavorare con dataset di dimensioni sempre maggiori, si sono sviluppate nuove tecniche che hanno di gran lunga migliorato le performance dei *clustering gerarchici*: gli algoritmi *clustering using representatives* (*CURE*), *robust clustering using links* (*ROCK*), *balanced iterative reducing and clustering using hierarchies* (*BIRCH*) (Xu et al., 2005).

Tutti i metodi gerarchici esaminati fino a questo punto sono metodi di tipo aggregativo, che partono dalla partizione banale dove ogni unità costituisce un cluster, per arrivare alla partizione che comprende ogni osservazione; come si è visto nell'introduzione sui vari metodi di *clustering* esistono anche dei metodi gerarchici di tipo divisivo, anche se meno utilizzati rispetto agli agglomerativi. I metodi gerarchici divisivi sono strutturati in maniera opposta a quelli agglomerativi, cioè consentono di ottenere una famiglia di partizioni delle n unità statistiche partendo da quella in cui tutte le unità sono riunite, per arrivare a quella in cui tutte sono divise. Tra questi si trovano gli algoritmi *divisive analysis* (*DIANA*) e *monothetic analysis* (*MONA*) (Xu et al., 2005).

3.2 I METODI NON GERARCHICI

I metodi di classificazione non gerarchici consentono di ottenere un'unica partizione degli n elementi in g gruppi ($g < n$): l'obiettivo è trovare una classificazione che soddisfi determinati criteri e che sia formata da un numero di gruppi g fissato a priori dal ricercatore. Solitamente tali metodi prevedono la specificazione esplicita di una funzione obiettivo, che viene spesso espressa in termini di scomposizione della devianza. Allora il processo di classificazione diventa un problema di ottimizzazione, dove si ricerca la partizione con la maggior omogeneità nei gruppi: questo significa che, almeno teoricamente, è possibile formalizzare il meccanismo di allocazione delle unità ai gruppi. Questo non è l'unico vantaggio di questa tipologia di metodi di *clustering*. Nei metodi non gerarchici viene meno il vincolo che tutte le coppie di unità che risultano tra loro unite ad un determinato livello di aggregazione gerarchica non possono più essere separate ai livelli successivi. Questo permette di superare i potenziali inconvenienti dovuti ad un raggruppamento errato nei passi iniziali di una procedura gerarchica.

Anche per questi metodi è possibile individuare una procedura generale che si può sintetizzare nelle seguenti fasi.

Fase 1: si sceglie una classificazione iniziale con un numero di cluster prefissato.

Fase 2: si calcola la variazione nella funzione obiettivo causata dallo spostamento di ciascuna unità dal cluster in cui si trova ad un altro e si sceglie per ciascuna unità il cluster che garantisce la maggiore omogeneità nei gruppi.

Fase 3: si ripete la fase 2 finché non viene verificata una regola di arresto prestabilita.

I metodi non gerarchici hanno quindi una struttura di tipo iterativo, che per un valore g prefissato rende l'algoritmo veloce; inoltre non è necessario costruire la matrice delle distanze. Questi vantaggi rendono questi metodi ideali nel caso di un dataset con un numero elevato di unità statistiche e nei casi nei quali lo studio vuole evidenziare le caratteristiche dei gruppi e non delle singole unità.

Il criterio dell'errore quadratico è la tecnica di *clustering* non gerarchica più intuitiva e più utilizzata. In particolare il metodo delle K-medie è il più semplice algoritmo che sfrutta questo criterio ed è implementato nei principali *packages* statistici.

L'algoritmo è caratterizzato da una procedura iterativa che consiste nei seguenti passi.

Fase 1. Si scelgono g "poli" (detti anche semi, o punti origine) iniziali, dei punti nello spazio p-dimensionale che costituiscono i centroidi dei gruppi della partizione iniziale. Questi poli possono essere individuati tramite metodi differenti, generalmente in maniera tale che siano sufficientemente distanti l'uno dall'altro. Quindi si ripartiscono le unità statistiche allocando ciascuna di esse al cluster il cui polo risulta più vicino, costituendo una partizione iniziale formata da g gruppi.

Fase 2. Si calcola per ogni unità la distanza dai centroidi di tutti i g cluster e ogni unità viene assegnata al cluster del centroide più vicino, qualora non vi fosse già allocata. In caso di riallocazione di un'unità si ricalcola il centroide sia del nuovo che del vecchio gruppo di appartenenza.

Fase 3. Si ripete la fase 2 fino a che l'algoritmo converge, cioè fino a quando non si verifica alcuna riallocazione rispetto all'iterazione precedente.

Le fasi della metodologia illustrata prevedono il calcolo ripetuto della distanza tra ogni unità ed i centroidi dei g cluster[4], per tale operazione viene solitamente utilizzata la distanza euclidea, in quanto essa garantisce la convergenza della procedura iterativa (Zani et al. (2007).

Ai fini di una corretta interpretazione dei risultati forniti da questo metodo è particolarmente influente la scelta della partizione iniziale.

Per quanto riguarda la scelta di g, il criterio probabilmente più diffuso consiste nell'esecuzione ripetuta dell'analisi con differenti valori di g e nella selezione della partizione più soddisfacente. Invece, per la scelta dei poli della partizione iniziale, un criterio molto semplice e poco dispendioso suggerisce di prendere le prime g osservazioni dell'insieme dei dati; un'altra regola prevede l'estrazione di un campione casuale delle unità.

Recentemente è stata sviluppata una nuova tecnica denominata *iterative self-organizing data analysis technique* (*ISODATA*)[5] che aggiusta in maniera dinamica il numero di cluster attraverso un *merging and splitting* dei cluster stessi, tenendo conto di una soglia prefissata. Dopo la prima fase, il nuovo K viene utilizzato come numero di cluster stimato per l'iterazione successiva. In questa categoria rientrano anche il *genetic K-means algorithm* (*GKA*) ed il *partitioning around medoids* (*PAM*).

4 Tale procedura appartiene alla classe di algoritmi di classificazione che adottano la tecnica denominata "ordinamento rispetto al centroide più vicino" (*nearest centroid sorting)*.

5 Per approfondimenti sugli algoritmi citati in questo paragrafo riguardante i metodi non gerarchici, dove non specificato diversamente, si veda Xu et al., 2005.

3.3 La *two step cluster analysis*

La *two step cluster analysis* inserita nel linguaggio di programmazione SPSS (Chiu et al., 2001) prevede due fasi:

Fase 1: *pre-clustering* delle unità.
In questa fase si utilizza un approccio sequenziale per costruire dei pre-cluster delle unità. Lo scopo è calcolare una nuova matrice di dimensioni minori da utilizzare nelle fasi successive; le unità che compongono tale matrice sono i pre-cluster, che vengono definiti come regioni dense dello spazio.
Il risultato può dipendere dall'ordine in cui sono organizzate le unità nella matrice di dati iniziale, conviene quindi usare un ordinamento casuale delle unità. Può essere utile ottenere più soluzioni con le unità estratte casualmente per verificare la stabilità di una soluzione specifica. Nei casi in cui questa operazione è complessa a causa delle dimensioni eccessive dei file, è possibile lavorare su campioni di unità.

Fase 2: *clustering* delle unità.
Si applica una tecnica gerarchica *model-based*. Analogamente agli algoritmi gerarchici agglomerativi, i pre-cluster sono uniti passo dopo passo fino ad arrivare alla fine del processo ad un unico cluster. A differenza degli algoritmi gerarchici classici si utilizza un modello che presuppone che le variabili quantitative continue si distribuiscano all'interno dei cluster come variabili normali indipendenti, mentre le variabili categoriali come variabili multinomiali indipendenti. Le due misure di prossimità disponibili sono rispettivamente: la distanza euclidea e la log-verosimiglianza in caso di variabili di tipo misto (Ming-Yi, 2010; Bacher et al., 2004).

3.4 Altri metodi

Alcuni metodi si basano sui modelli mistura. L'assunzione di fondo di queste tipologie di *clustering* è che i dati siano generati da diverse distribuzioni dello stesso tipo e l'obiettivo è di identificare i parametri di ognuna e il loro numero; solitamente si assume che le componenti individuali della densità mistura siano variabili normalmente distribuite. La definizione di questi metodi si basa sul concetto di modello mistura: per questo si utilizza il termine *model-based clustering*. L'approccio più tradizionale prevede di ottenere, in maniera iterativa, una stima di massima verosimiglianza dei vettori dei parametri delle densità; più recentemente viene utilizzato l'algoritmo *EM* (*Expectation Maximization*) (McLachlan et al., 2000).

Altri metodi utilizzano la teoria dei grafi. Questa costituisce uno strumento molto potente, in quanto, pur nella relativa semplicità della nozione di grafo, consente di descrivere in maniera strutturata e formalizzata matematicamente problemi e situazioni correnti di una certa complessità. In questo caso i nodi del grafo corrispondono alle unità e gli archi rappresentano la prossimità tra ogni coppia di punti.

Il *clustering identification via connectivity kernels* (*CLICK*), l'algoritmo più conosciuto in questo ambito, prevede come prima fase la costruzione del *minimal spanning tree* (*MST*) dei dati; quindi si eliminano gli archi di maggior lunghezza per generare i cluster. Un algoritmo sviluppatosi recentemente, ma molto utilizzato, è *Chameleon*: questo algoritmo procede all'eliminazione di un arco se entrambi i vertici non sono compresi tra i punti più vicini relativi ad entrambi. Altri algoritmi utilizzati sono *Delaunay triangulation graph* (*DTG*), *highly connected subgraphs* (*HCS*), *cluster affinity search technique* (*CAST*).

Recentemente si sono sviluppati nuovi metodi di *clustering*, che trovano fondamento in teorie non utilizzate solitamente per queste applicazioni.

Molto utilizzati in ambito spaziale sono gli algoritmi basati sulla densità: questi considerano i cluster come regioni di spazio ad alta densità, separate tra loro da regioni a bassa densità. Gli algoritmi basati

sulla densità analizzano la densità attorno ad ogni osservazione e la classificano come "sufficientemente densa" se il numero di osservazioni prese è maggiore rispetto una certa soglia prefissata. L'algoritmo più utilizzato è il *density based spatial clustering of applications with noise* (*DBSCAN*), che separa in diversi cluster le regioni con densità sufficientemente elevata: questo avviene andando ad osservare per ogni punto lo spazio determinato da un raggio fissato, per vedere se i punti presenti in tale area superano il numero imposto a priori come soglia. Un altro metodo è il *density-based clustering* (*DENCLUE*): questo algoritmo si basa sull'idea che ogni punto possa essere modellato utilizzando una funzione matematica chiamata *influence function*, che descrive l'impatto dell'osservazione sui punti vicini. Sommando queste funzioni è possibile trovare la densità dello spazio dei punti ed i cluster possono essere trovati determinando matematicamente i massimi locali della funzione di densità così costruita (Han et al., 2001).

Vi sono metodi basati sulle tecniche di ricerca combinatoria, che vedono il *clustering* come una sorta di problema di ottimizzazione: queste tecniche hanno come obiettivo principale la ricerca dell'ottimo (o di una sua approssimazione) in un problema di ottimizzazione combinatoria. Considerato un dataset di punti $\boldsymbol{x}_j \in \mathbb{R}^d, j = 1, \dots, N$, l'algoritmo di *clustering* ha lo scopo di organizzare questi punti in K gruppi $\{C_1, \dots, C_K\}$ in modo da ottimizzare una qualche funzione obiettivo. Con le tecniche di ricerca utilizzate solitamente nella ricerca comnbinatoria non viene garantita l'ottimalità della partizione ottenuta, per cui si ricorre a metodi di ricerca più complessi, quali i *genetically guided algorithm* (*GGA*), il *tabu search* (*TS*) *clustering* e il *deterministic annealing* (*DA*) *clustering* (Xu, 2005).

Esistono anche metodi basati sulle reti neurali, modelli matematici/ informatici di calcolo basati sulle reti neurali biologiche. Gli algoritmi più utilizzati in questo campo sono: *learning vector quantization* (*LVQ*) e *self-organizing feature map* (*SOFM*).

4. L'organizzazione dei dati della matrice del pendolarismo 2011

4.1 La matrice del pendolarismo

La matrice del pendolarismo contiene informazioni relative ai movimenti per studio e per lavoro dei pendolari, ossia di quella parte della popolazione residente che ha dichiarato di recarsi "giornalmente al luogo di studio o lavoro partendo dall'alloggio di residenza e di rientrare giornalmente allo stesso"[6]. Ogni *record* della matrice si riferisce ad un gruppo di pendolari (strato) che presentano le stesse caratteristiche (comune di residenza, comune di destinazione, motivo dello spostamento, sesso etc.). La prima matrice costruita dall'Istat si riferisce al 1981 e riporta informazioni raccolte in occasione del Censimento; le successive sono state costruite con i dati dei Censimenti del 1991 e del 2001. In seguito alle innovazioni introdotte in occasione del Censimento 2011 (Borrelli et al., 2012) e all'introduzione del campionamento nei comuni al di sopra dei 20000 abitanti per la raccolta delle informazioni relative alle variabili socio-economiche, tra cui anche parte di quelle relative al pendolarismo, la matrice riferita al 2011 presenta caratteristiche diverse rispetto alle precedenti. I *record* che la compongono sono infatti di due tipi, individuati dalla lettera S (che richiama *short*) e la lettera L (che richiama *long*). Le informazioni censuarie sono state infatti raccolte usando due tipi di questionari, uno in forma estesa (*long*) che è stato distribuito nei comuni fino a 20000 abitanti e ad un campione di famiglie nei comuni capoluogo di provincia o con almeno 20000 abitanti, ed un questionario in forma ridotta (*short*) distribuito alle restanti famiglie di questi ultimi comuni.

I *record* che compongono la matrice 2011 si differenziano per il fatto che nei *record* di tipo L le variabili di stratificazione sono in numero maggiore infatti oltre a quelle riportate per il tipo di *record*

6 Si veda http://www.istat.it/it/archivio/139381.

S (provincia di residenza, comune di residenza, sesso, motivo dello spostamento, luogo di studio o di lavoro, provincia abituale di studio o lavoro, comune abituale di studio o lavoro, stato estero di studio o lavoro) ve ne sono altre tre (mezzo, orario di uscita, tempo impiegato). Queste informazioni sono state rilevate solo con il questionario di tipo *long*. La matrice contiene inoltre due ulteriori variabili, che rappresentano i pesi degli strati; una di esse fa riferimento ai *record* di tipo S ed in questo caso il peso è il numero di pendolari appartenenti a quello strato, l'altra invece fa riferimento ai *record* di tipo L dove i pesi associati a ciascuno stato sono delle stime, dal momento che le maggiori informazioni contenute in tali strati (mezzo di trasporto, orario, tempo utilizzato) sono state raccolte su base campionaria nei comuni con almeno 20000.

4.2 La descrizione e l'organizzazione dei dati utilizzati

Nel presente lavoro si sono utilizzate le informazioni relative ai pendolari residenti in Friuli Venezia Giulia. Dalla matrice nazionale costruita per tutti i comuni italiani è stata quindi estratta la matrice per il Friuli Venezia Giulia. Essa contiene 134293 *record* (suddivisi in 28126 di tipo S e 106167 di tipo L), che raccolgono informazioni relative a 617439 persone residenti in regione (in famiglia o in convivenza[7]) che dichiarano di muoversi giornalmente verso il luogo di studio o lavoro e di rientrare la sera all'abitazione da cui sono partiti. Per le sole persone residenti in famiglia (616993) nei *record* di tipo L (106167) sono riportate anche le informazioni relative ai mezzi, orari e tempi.

Nel lavoro sono stati usati i soli *record* di tipo L, dal momento che essi contengono le informazioni relative a mezzo utilizzato, orario di uscita dall'abitazione e tempo impiegato per raggiungere il luogo di studio o lavoro. Tali *record* fanno riferimento ai soli pendolari che vivono in famiglia e che rappresentano nel caso del Friuli Venezia Giulia il 99,9% dei pendolari della regione.

Come già accennato ogni *record* rappresenta una tipologia di pendolare individuata sulla base delle variabili descritte nella Tavola 4. L'ultima variabile riportata nella Tavola 4 (Stima del numero di pendolari) rappresenta il peso di ciascuno strato di tipo L.

È opportuno precisare che durante il Censimento viene rilevata una sola modalità di trasporto, anche nel caso in cui il pendolare utilizzi più modalità, e precisamente quella adottata per compiere il tragitto più lungo, in termini di distanza percorsa. Inoltre il tempo per arrivare al luogo di lavoro comprende anche il tempo utilizzato per eventuali tragitti più lunghi, come ad esempio per portare i bambini a scuola.

Poiché l'oggetto dello studio è il pendolare, è stato necessario espandere la matrice in modo tale che ad ogni *record* corrispondesse un pendolare, e non più uno strato. In tal modo si sono ottenuti tanti *record* quanti sono i pendolari residenti in famiglia e si è potuto procedere all'applicazione della *two step cluster analysis*, che permette di considerare sia le variabili qualitative che quantitative presenti nella matrice.

Per l'applicazione sono state utilizzate tre variabili qualitative e due quantitative. Le variabili qualitative sono: motivo dello spostamento, luogo di lavoro o studio, modalità di trasporto. Su queste variabili non è stata effettuata alcuna trasformazione e le loro modalità sono riportate nella Tavola 4.

---

7 Insieme di persone che, senza essere legate da vincoli di matrimonio, parentela, affinità e simili, conducono vita in comune per motivi religiosi, di cura, di assistenza, militari, di pena e simili. Le persone addette alla convivenza per ragioni di lavoro, se vi convivono abitualmente, sono considerate membri permanenti della convivenza purché non costituiscano famiglia a sé stante. I principali tipi di convivenza sono: istituti d'istruzione, istituti assistenziali, istituti di cura pubblici e privati, istituti penitenziari, convivenze ecclesiastiche, convivenze militari e di altri corpi accasermati, alberghi, pensioni, locande e simili, navi mercantili, altre convivenze (ad esempio, case dello studente), da http://www3.istat.it/cgi-bin/glossario/indice.pl.

Tavola 4 – Variabili della matrice del pendolarismo al Censimento 2011 (*record* L)

	Tipo di variabile	Modalità della variabile
Provincia di residenza	Qualitativa	codice Istat
Comune di residenza	Qualitativa	codice Istat
Sesso	Qualitativa	maschio, femmina
Motivo dello spostamento	Qualitativa	lavoro, studio
Luogo di lavoro o studio	Qualitativa	comune di resid., altro comune, estero
Provincia di lavoro o studio	Qualitativa	codice Istat
Comune di lavoro o studio	Qualitativa	codice Istat
Paese estero di lavoro o studio	Qualitativa	codice Istat
Modalità di trasporto	Qualitativa	treno; tram; autobus urbano o filobus; corriera o autobus extra-urbano; autobus aziendale o scolastico; auto privata (come conducente); auto privata (come passeggero); motocicletta o ciclomotore o scooter; bicicletta; altro mezzo; a piedi;
Orario di uscita dall'abitazione	Quantitativa	prima delle 7:15; dalle 7:15 alle 8:14; dalle 8:15 alle 9:14; dopo le 9:14;
Tempo per arrivare al luogo di lavoro o studio	Quantitativa	fino a 15 minuti; da 16 a 30 minuti; da 31 a 60 minuti; oltre 60 minuti;
Stima del numero di pendolari	Quantitativa	

Le due variabili quantitative, derivate invece da una trasformazione delle due variabili originarie (orario di uscita dall'abitazione e tempo impiegato), sono:

– @orario: sono stati considerati quattro valori corrispondenti ai valori centrali delle classi della variabile originaria, considerate tutte della stessa ampiezza. Solo la prima classe ("prima delle 7:45") è stata considerate più ampia per tener conto dei pendolari che partono molto presto al mattino.
– @tempo: I valori della variabile corrispondono ai valori centrali della variabile originaria. La classe "oltre 60 minuti" è stata considerata della stessa ampiezza della precedente.

L'analisi è stata condotta separatamente per i comuni con almeno 20000 abitanti (Gorizia, Pordenone, Trieste, Udine, Monfalcone e Sacile) e per quelli sotto i 20000. La distinzione tra questi due "tipi di comune" è stata mantenuta anche in considerazione delle diverse modalità di raccolta delle informazioni.

In tutti i casi il numero dei cluster è stato scelto sulla base del *Bayesian Information Criteria* (*BIC*).

Tavola 5 – Popolazione residente e pendolari nei comuni del Friuli Venezia Giulia al Censimento 2011

Comuni	Popolazione residente (R)	Pendolari residenti in famiglia (P*)		P*/R %
		per studio	per lavoro	
Gorizia	35212	4760	11601	46,5
Pordenone	50583	7105	17198	48,0
Trieste	202123	33319	70392	40,1
Udine	98287	77831	34142	48,8
Monfalcone	27041	3604	8282	44,0
Sacile	19897	3702	7180	54,7
Comuni <20000 abitanti	785842	116814	286708	51,3
Totale FVG	1218985	178584	438409	50,6

Fonte: Istat – Censimento 2011

5. Il pendolarismo nei comuni sotto i 20000 abitanti

Nel caso dei pendolari per motivo di studio nei comuni sotto i 20000 abitanti si sono evidenziati due cluster, al primo appartiene il 44,3% dei pendolari, al secondo il rimanente 55,7% (si veda Tavola 6). Per quanto riguarda la distribuzione secondo sesso non vi sono differenze significative tra i due cluster; vi è invece una sostanziale differenza per quanto riguarda l'ora di partenza da casa (@orario) e il tempo di arrivo al lavoro (@tempo) e le modalita di trasporto.

In dettaglio:

Cluster 1 – è formato da individui che studiano in un comune diverso da quello di residenza, escono da casa presto (in media verso le 7) e impiegano in media circa trenta minuti per raggiungere il luogo di studio utilizzando prevalentemente una corriera o un'autobus, ma anche l'auto come passeggero.

Cluster 2 – vi appartengono pendolari che studiano nel comune di residenza, escono da casa circa un'ora dopo coloro che appartengono al Cluster 1 e impiegano poco più di 10 minuti per raggiungere il luogo di studio, prevalentemente come passeggeri di un'auto.

Considerando i pendolari per motivi di lavoro si sono individuati cinque cluster. La distribuzione dei pendolari è presentata nella Tavola 6. Il cluster più numeroso è il Cluster 5 seguito dal Cluster 4.

Tavola 6 – Distribuzione dei pendolari nei cluster secondo motivo dello spostamento

	Studio			Lavoro	
	Numerosità	**Percentuale**		**Numerosità**	**Percentuale**
Cluster 1	51779	44,3%	Cluster 1	48012	16,7%
Cluster 2	65035	55,7%	Cluster 2	28373	9,9%
			Cluster 3	37932	13,2%
			Cluster 4	76649	26,7%
			Cluster 5	95742	33,4%
Totale	116814	100,0%	Totale	286708	100,0%

Nel Cluster 1 e nel Cluster 5 sono presenti solo individui di sesso maschile mentre nei cluster Cluster 3 e nel Cluster 4 ci sono solo rappresentanti del sesso femminile.

Le modalità di trasporto differiscono nei cinque cluster. Ad esempio i pendolari appartenenti ai Cluster 4 e Cluster 5 utilizzano esclusivamente l'autovettura come guidatore, mentre i pendolari del Cluster 2 usano tutte le altre modalità a parte una piccola parte che usa la macchina. Nel dettaglio i profili dei pendolari per motivi di lavoro nei diversi cluster è:

Cluster 1 – composto da maschi che lavorano nello stesso comune di residenza utilizzano per la maggior parte la loro macchina (62%), ma si muovono anche a piedi o in bicicletta (32%), escono di casa dopo le 7 di mattina ed impiegano circa 10 minuti per raggiungere il posto di lavoro.

Cluster 3 – composto da femmine che lavorano nello stesso comune di residenza che escono da casa dopo gli appartenenti al Cluster 1 utilizzando modalità di trasporto simili.

Cluster 4 – composto da femmine che lavorano in communi diversi da quello di residenza, utilizzano la macchina escono da casa prima delle donne del Cluster 3 ed impiegano in media circa venti minuti per raggiungere il posto di lavoro.

Cluster 5 – è composto da maschi che escono da casa per andare al lavoro prima delle femmine del Cluster 4, dal quale non differiscono né per il mezzo utilizzato né per il tempo impiegato.

Cluster 2 – è composto da maschi e femmine che lavorano principalmente fuori del comune di residenza, tendono a non utilizzare la macchina come guidatore ed utilizzano invece tutti gli altri mezzi (compresa l'auto privata come passeggero). Escono da casa presto, in genere prima delle persone appartenenti agli altri cluster ed hanno i tempi di percorrenza più lunghi (in media 30 minuti). Si tratta del cluster meno numeroso.

6. I PENDOLARI DEI COMUNI CON ALMENO 20000 ABITANTI E DEI COMUNI CAPOLUOGO

La Tavola 7 riporta la distribuzione percentuale dei pendolari nei cluster secondo motivo dello spostamento nei comuni di maggiori dimensioni. Sia nel caso dello spostamento per motivi di studio che di lavoro si sono individuati da due a tre cluster all'interno di ciascun comune, a parte il caso del comune di Pordenone che presenta quattro cluster di pendolari per motivi di studio.

Tavola 7 – Distribuzione percentuale dei pendolari nei cluster secondo comune di residenza e motivo dello spostamento

	Studio			**Lavoro**	
Gorizia	Cluster 1	13,0%	**Gorizia**	Cluster 1	34,7%
	Cluster 2	39,3%		Cluster 2	65,3%
	Cluster 3	47,7%			
		100,0%			100,0%
Pordenone	Cluster 1	10,0%	**Pordenone**	Cluster 1	43,6%
	Cluster 2	30,6%		Cluster 2	28,8%
	Cluster 3	23,2%		Cluster 3	27,7%
	Cluster 4	36,2%			
		100,0%			100,0%
Trieste	Cluster 1	37,6%	**Trieste**	Cluster 1	27,6%
	Cluster 2	33,1%		Cluster 2	48,3%
	Cluster 3	29,3%		Cluster 3	24,1%
		100,0%			100,0%
Udine	Cluster 1	25,1%	**Udine**	Cluster 1	34,5%
	Cluster 2	36,2%		Cluster 2	65,5%
	Cluster 3	38,7%			
		100,0%			100,0%
Monfalcone	Cluster 1	28,6%	**Monfalcone**	Cluster 1	46,8%
	Cluster 2	71,4%		Cluster 2	33,1%
				Cluster 3	20,1%
		100,0%			100,0%
Sacile	Cluster 1	22,3%	**Sacile**	Cluster 1	59,6%
	Cluster 2	42,8%		Cluster 2	40,4%
	Cluster 3	34,9%			
		100,0%			100,0%

Nelle Tavole dalla 8 alla 13 vengono riportati per ciascun comune i valori medi e i valori modali delle variabili utilizzate nella *cluster analysis*.

Tavola 8 – Descrizione dei cluster nel comune di Gorizia

	Orario di uscita dall'abitazione in media (ore e minuti)	Tempo utilizzato in media (minuti)	Sesso		Modalità di spostamento più frequente	Luogo di studio o lavoro prevalente
			M	F		
Motivi di studio						
Cluster 1	7:20	49	50%	50%	Treno (61%)	Altro comune (88%)
Cluster 2	8:00	11	53%	47%	A piedi (65%)	Comune di residenza (100%)
Cluster 3	7:51	9	50%	50%	Auto (come passeggero) (100%)	Comune di residenza (100%)
Motivi di lavoro						
Cluster 1	7:29	30	60%	40%	Auto (come conducente) (70%)	Altro comune (83%)
Cluster 2	7:38	17	50%	50%	Auto (come conducente) (65%)	Comune di residenza (100%)

Tavola 9 – Descrizione dei cluster nel comune di Pordenone

	Orario di uscita dall'abitazione in media (ore e minuti)	Tempo utilizzato in media (minuti)	Sesso		Modalità di spostamento più frequente	Luogo di studio o lavoro prevalente
			M	F		
Motivi di studio						
Cluster 1	7:29	45	44%	56%	Treno (42%)	Altro comune (94%)
Cluster 2	7:38	14	55%	45%	Bicicletta (62%)	Comune di residenza (100%)
Cluster 3	7:56	10	48%	52%	A piedi (100%)	Comune di residenza (100%)
Cluster 4	7:55	9	52%	48%	Auto (come passeggero) (100%)	Comune di residenza (100%)
Motivi di lavoro						
Cluster 1	7:27	20	97%	3%	Auto (come conducente) (87%)	Altro comune (62%)
Cluster 2	7:46	16	-	100%	Auto (come conducente) (100%)	Comune di residenza (56%)
Cluster 3	7:54	11	40%	60%	Bicicletta (44%)	Comune di residenza (94%)

Tavola 10 – Descrizione dei cluster nel comune di Trieste

	Orario di uscita dall'abitazione in media (ore e minuti)	Tempo utilizzato in media (minuti)	Sesso		Modalità di spostamento più frequente	Luogo di studio o lavoro prevalente
			M	F		
Motivi di studio						
Cluster 1	8:00	12	53%	47%	Auto (come passeggero) (72%)	Comune di residenza (100%)
Cluster 2	7:30	26	48%	52%	Autobus urbano (91%)	Comune di residenza (93%)
Cluster 3	8:00	9	52%	48%	A piedi (100%)	Comune di residenza (100%)
Motivi di lavoro						
Cluster 1	7:27	20	97%	3%	Auto (come conducente) (87%)	Altro comune (62%)
Cluster 2	7:46	16	-	100%	Auto (come conducente) (100%)	Comune di residenza (56%)
Cluster 3	7:54	11	40%	60%	Bicicletta (44%)	Comune di residenza (94%)

Tavola 11 – Descrizione dei cluster nel comune di Udine

	Orario di uscita dall'abitazione in media (ore e minuti)	Tempo utilizzato in media (minuti)	Sesso		Modalità di spostamento più frequente	Luogo di studio o lavoro prevalente
			M	F		
Motivi di studio						
Cluster 1	7:12	25	47%	53%	Autobus urbano (70%)	Comune di residenza (78%)
Cluster 2	7:56	10	51%	49%	Auto (come passeggero) (100%)	Comune di residenza (100%)
Cluster 3	8:03	10	52%	48%	A piedi (58%)	Comune di residenza (100%)
Motivi di lavoro						
Cluster 1	7:05	26	63%	37%	Auto (come conducente) (83%)	Altro comune (100%)
Cluster 2	7:46	14	46%	54%	Auto (come conducente) (54%)	Comune di residenza (100%)

Tavola 12 – Descrizione dei cluster nel comune di Monfalcone

	Orario di uscita dall'abitazione in media (ore e minuti)	Tempo utilizzato in media (minuti)	Sesso		Modalità di spostamento più frequente	Luogo di studio o lavoro prevalente
			M	F		
Motivi di studio						
Cluster 1	7:11	37	55%	45%	Corriera, autobus extra-urbano (30%)	Altro comune (98%)
Cluster 2	7:44	9	53%	47%	A piedi (42%)	Comune di residenza (39%)
Motivi di lavoro						
Cluster 1	7:29	27	58%	42%	Auto (come conducente) (67%)	Altro comune (86%)
Cluster 2	7:14	11	100%	-	Auto (come conducente) (53%)	Comune di residenza (100%)
Cluster 3	7:52	10	-	100%	Auto (come conducente) (62%)	Comune di residenza (100%)

Tavola 13 – Descrizione dei cluster nel comune di Sacile

	Orario di uscita dall'abitazione in media (ore e minuti)	Tempo utilizzato in media (minuti)	Sesso		Modalità di spostamento più frequente	Luogo di studio o lavoro prevalente
			M	F		
Motivi di studio						
Cluster 1	6:17	41	54%	46%	Corriera, autobus extra-urbano (46%)	Altro comune (100%)
Cluster 2	7:53	10	51%	49%	Auto (come passeggero) (100%)	Comune di residenza (79%)
Cluster 3	7:43	13	52%	48%	A piedi (45%)	Comune di residenza (100%)
Motivi di lavoro						
Cluster 1	7:33	22	58%	42%	Auto (come conducente) (88%)	Altro comune (100%)
Cluster 2	7:41	10	51%	49%	Auto (come conducente) (61%)	Comune di residenza (100%)

Per quanto riguarda il pendolarismo per motivi di studio a Gorizia vi sono due cluster (Cluster 2 e Cluster 3) composti da persone che studiano nello stesso comune di residenza, la differenza consiste nel fatto che nel Cluster 2 vanno prevalentemente a piedi, mentre nel Cluster 3 in auto (come passeggeri). Il Cluster 1 è composto in prevalenza da individui che studiano in un altro comune e utilizzano in maggioranza il treno. Analoghe considerazioni si possono fare per il comune di Pordenone; in questo però vengono identificati quattro cluster. Ad uno di essi (Cluster 2) appartengono pendolari che utilizzano prevalentemente la bicicletta. Nei comuni di Trieste e di Udine si studia prevalentemente nel comune di residenza e tra i cluster individuati vi è uno in cui il mezzo di trasporto principale è l'autobus urbano (Cluster 2 a Trieste e Cluster 1 a Udine). Per quanto riguarda i due comuni minori non capoluogo di provincia (Monfalcone e Sacile) essi presentano situazioni simili, con l'utilizzo prevalente nel Cluster 1 dell'autobus extra-urbano. A Sacile è stato isolato un ulteriore cluster composto totalmente da studenti che utilizzano l'auto (come passeggero) per spostamenti all'interno del comune di residenza, come accade nella maggioranza dei piccoli comuni.

I cluster individuati per motivi di lavoro nei vari comuni sono formati in prevalenza da pendolari che utilizzano l'auto come conducente. Solo a Pordenone e a Trieste vi sono dei cluster in cui le modalità di spostamento prevalenti sono diverse: bici a Pordenone (Cluster 3) e a Trieste motocicletta (Cluster 3) e autobus urbano (Cluster 1). Tutti i comuni tranne Pordenone e Trieste presentano due cluster con caratteristiche simili (a Monfalcone i cluster diventano tre per effetto della variabile sesso).

7. Conclusioni

In questo lavoro viene utilizzata la matrice del pendolarismo derivata dal Censimento 2011 per individuare le tipologie di pendolari della regione Friuli Venezia Giulia. Ciascun *record* della matrice rappresenta uno strato di pendolari descritto dalle variabili presenti nella matrice stessa.

La metodologia utilizzata per l'analisi, la *two step cluster*, ha consentito di ottenere dei gruppi ben definiti di pendolari con riferimento al Friuli Venezia Giulia, e distinguendo i comuni sopra e sotto i 20000 e le motivazioni dello spostamento.

Nei comuni sotto i 20000 abitanti sono stati identificati due cluster di pendolari per motivi di studio e cinque cluster per motivi di lavoro. Le variabili utilizzate discriminano bene in entrambi i casi.

Nel caso invece dei comuni con almeno 20000 abitanti, in cui l'analisi è stata condotta singolarmente per ciascun comune, vengono individuati un numero di cluster per motivi di lavoro che va da due (Gori-

zia, Udine, Sacile) a tre (Monfalcone, Pordenone, Trieste); i cluster individuati per motivi di studio sono più numerosi a Pordenone (quattro) mentre in tutti gli altri comuni ne sono stati individuati tre, escluso Monfalcone in cui la suddivisione migliore ha portato a due cluster.

La *two step cluster analysis* utilizzata nel presente lavoro ha permesso di segmentare i pendolari arrivando ad individuare un numero decisamente ridotto di raggruppamenti (da 2 ad un massimo di 5). Tale segmentazione è utile per poter differenziare l'offerta di mezzi di trasporto pubblico e di servizi a supporto del trasporto privato, dal momento che è possibile conoscere per ogni comune quale è il segmento di pendolari prevalente.

Sulla base delle evidenze ottenute da questa analisi potrebbe essere utile sviluppare politiche di incentivo e di incremento del trasporto pubblico, che tengano conto delle diverse di tipologie di pendolari e di come esse si distribuiscono sul territorio regionale, oltre che politiche di mobilità sostenibile rivolte ad esempio all'incentivazione dell'uso dell'auto elettrica.

Un altro aspetto riguarda il possibile sviluppo di modalità innovative di trasporto, come ad esempio il *car sharing* e il *car pooling*. Con *car sharing* si intende un servizio per utilizzare un automobile su prenotazione prelevandola e riportandola in un parcheggio, mentre con il *car pooling* più persone viaggiano sulla stessa auto di proprietà di uno dei viaggiatori dividendo tra loro le spese. Questi servizi vengono cosi incontro non solo alla soddisfazione delle esigenze di flessibilità nella mobilità che solitamente viene ottenuta con l'utilizzo dell'auto propria, ma anche favoriscono il passaggio dal possesso del mezzo all'uso dello stesso.

Riferimenti bibliografici

BACHER, J., K. WENZING, M., VOGLER, M. (2004) "SPSS Two Cluster A First Evaluation", *Universitat Erlangen-Nurnberg*, pp. 1-20, www.statisticalinnovations.com/products/twostep.pdf cited July, 2014.

BORRELLI, F., CARBONETTI, G., DE FELICI, L., SOLARI, F. (2012) "Metodologie di stima per piccole aree applicabili a variabili di censimento", *Istat Working Papers* 3-2012, ISTAT, Roma.

CHIU, T., FANG, D., CHEN, J., WANG, Y., JERIS, C. (2001) "A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment", *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2001*, pp. 263-268.

GOODMAN, A. (2013) "Walking, Cycling and Driving to Work in the English and Welsh 2011 Census: Trends, Socio-Economic Patterning and Relevance to Travel Behaviour", *PLoS ONE* 8(8): e71790. doi: 10.1371/journal.pone.0071790.

HAN, J., KAMBER, M., TUNG, A.K.H. (2001) *Spatial clustering methods in data mining. A survey*, http://www.cs.uiuc.edu/homes/hanj/

MCLACHLAN, G.J., PEEL, D. (2000) *Finite Mixture Models*, Wiley, New York.

MING-YI, S., JAR-WEN, J., LIEN-FU, L. (2010) "A Two-Step Method for Clustering Mixed Categorical and Numeric Data", *Tamkang Journal of Science and Engineering* 13 (1), pp. 11-19.

STASSI, G., VALENTINI, A. (2013) *L'Italia del Censimento. Struttura demografica e processo di rilevazione. Friuli Venezia Giulia*, Istat.

ZANI, S., CERIOLI, A. (2007) *Analisi dei dati e data mining per le decisioni aziendali*, Giuffrè Editore, Milano.

XU, R., WUNSCH, D. (2005) "Survey of Clustering Algorithms", *IEEE transactions on neural networks* 16(3), pp. 645-678.